Exploratory Data Analysis of Spotify Dataset

ASMIT DEY

Spotify Tracks

Problem Statement Summary:

The Music Director/Mixing Engineer aiming to optimize new songs for popularity needs to leverage insights from Spotify tracks data. The core challenge is to understand audio features, trends, and patterns that drive track popularity to inform production and mixing decisions.

Important Points:

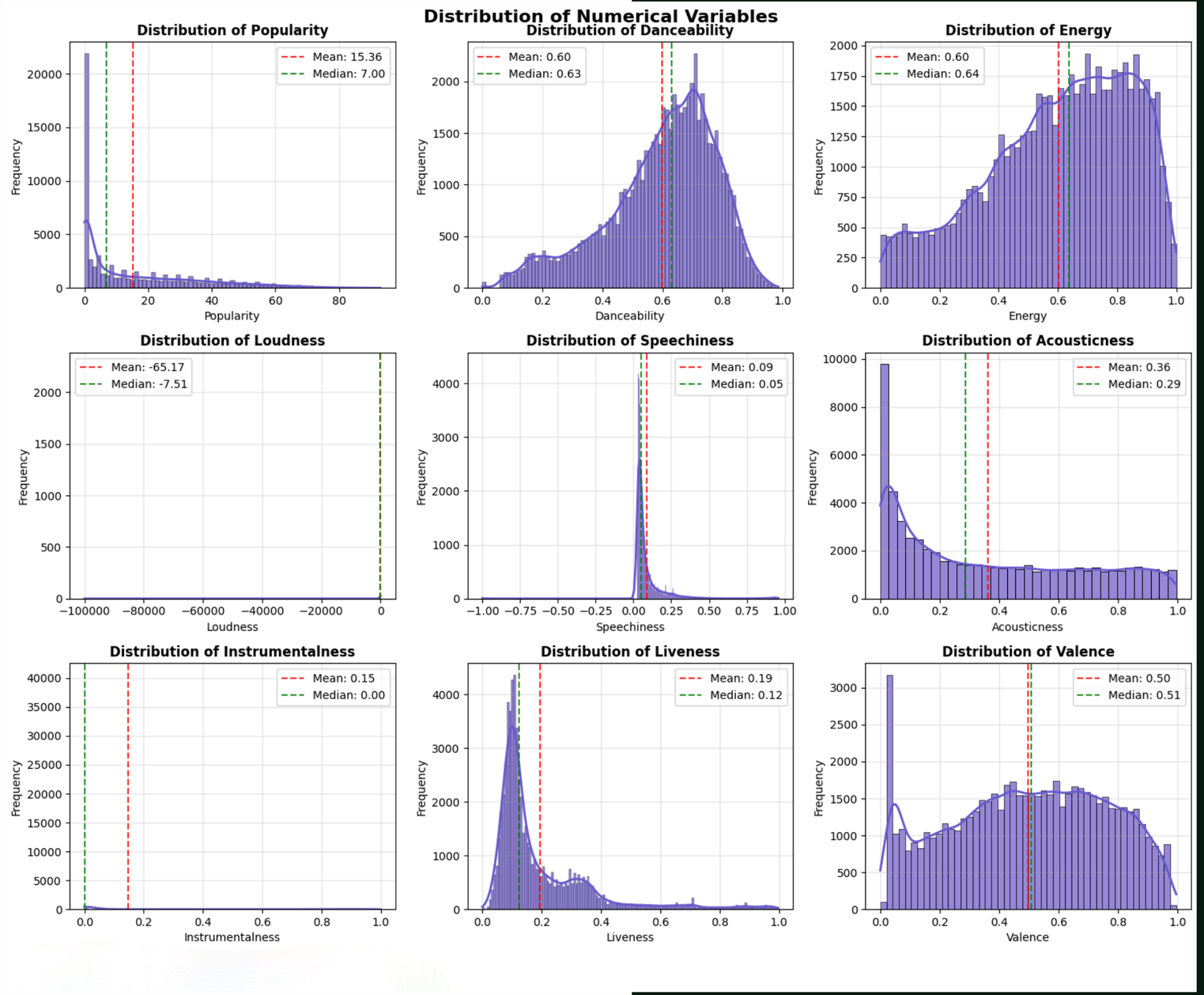
Owner/User: Music Director/Mixing Engineer.
Context: Collection of Spotify tracks with audio features and metadata.
Consumer of Analysis: The Music Director/Mixing Engineer themselves.

Expectation:

- Analysis: Deep dive into distributions, correlations, and trends in popularity, audio features (danceability, energy, valence, etc.), and metadata (year, language, key).
- Insight Identification: Uncover trends, top/bottom performing features/artists/languages, correlations with popularity, gaps, and yearly patterns.
- Strategic Recommendation: Provide concrete, actionable strategies to optimize song production and mixing for higher popularity.

Key Challenge Areas Revealed by Analysis:

- Optimizing for high-popularity sound profiles with significant emphasis on danceability, energy, and valence peaks.
- Maximizing the contribution of top-performing artists, languages, and feature combinations.
- Addressing the underperformance and potential inefficiency of certain audio characteristics (e.g., low speechiness or instrumentalness).
- Leveraging yearly trends by understanding shifts in popular features like loudness and duration.
- Improving mixes for modern standards in loudness, tempo, and mode.
- Replicating success factors from high-popularity tracks across new productions.

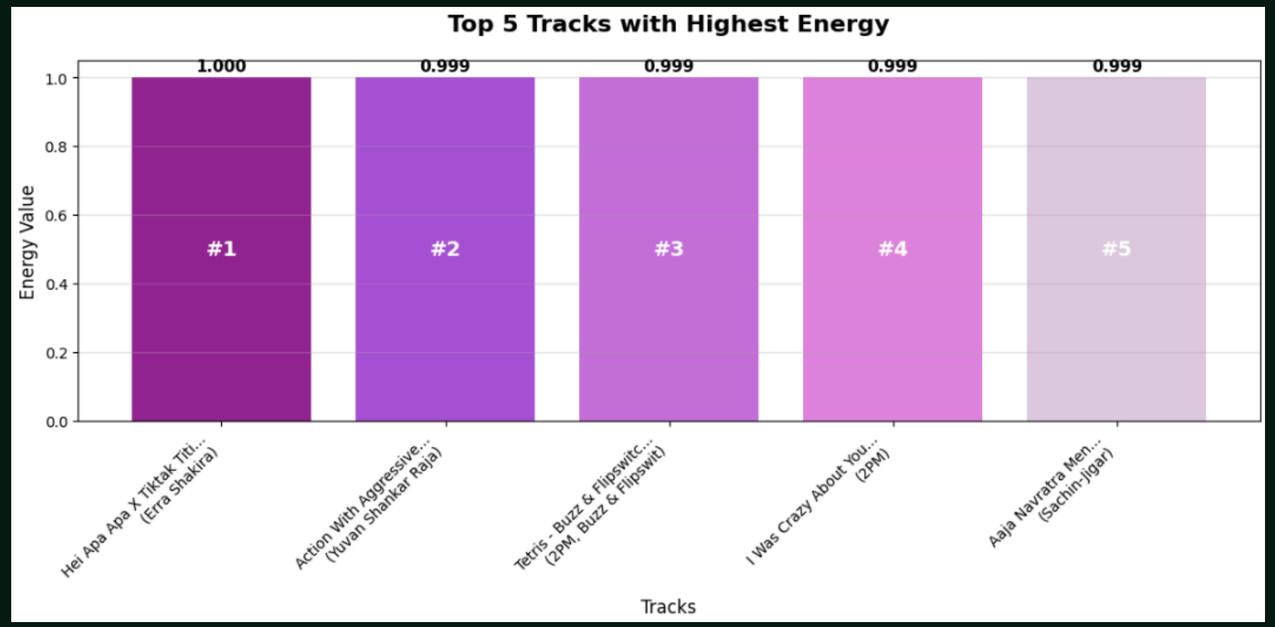


Visualization of Data Distributions using Histograms with KDE

Summary of the histogram plots :

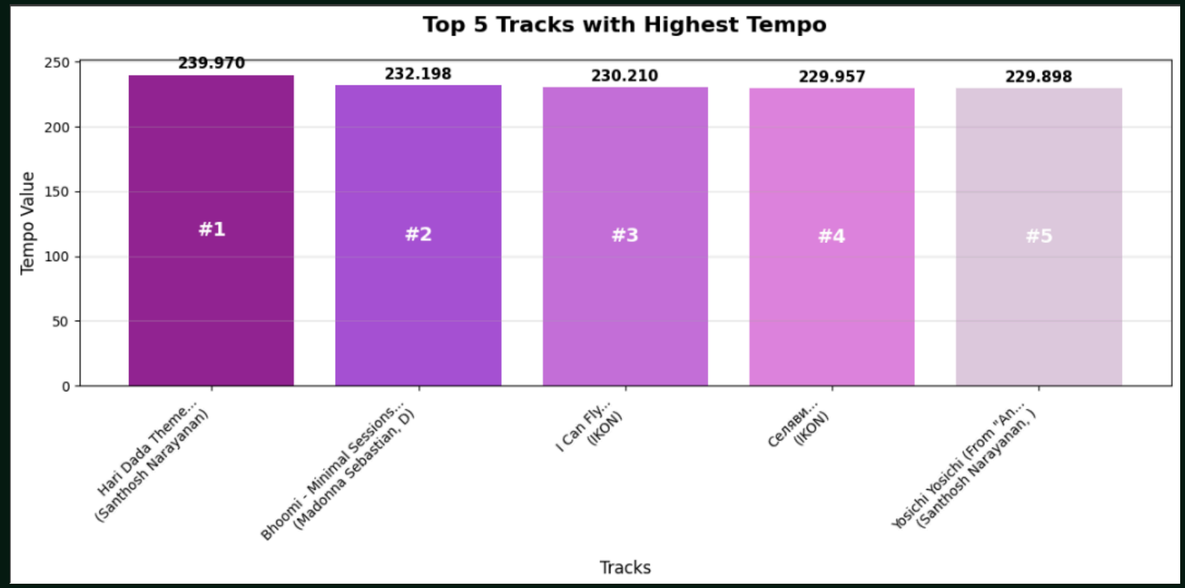
- Popularity: Right-skewed; most tracks have low popularity with few highly popular ones.
- Danceability & Energy: Both show moderate to high values, peaking around 0.6–0.8, suggesting most songs are upbeat and energetic.
- Loudness: Majority centered between -10 dB to -5 dB, but extreme negative outliers exist due to data errors (e.g., -100000).
- Speechiness: Mostly low (<0.1), indicating fewer spoken-word tracks.
- Acousticness: Broad spread; slight concentration toward lower values, meaning most songs are less acoustic.
- Instrumentalness: Highly right-skewed with most near zero — few purely instrumental tracks.
- Liveness: Mostly below 0.3, suggesting limited live-recorded tracks.
- Valence: Fairly uniform, showing balanced emotional tones across tracks.
- Tempo: Roughly normal distribution centered around 118 BPM; a few invalid negatives present.

Bar Chart With Highest Energy



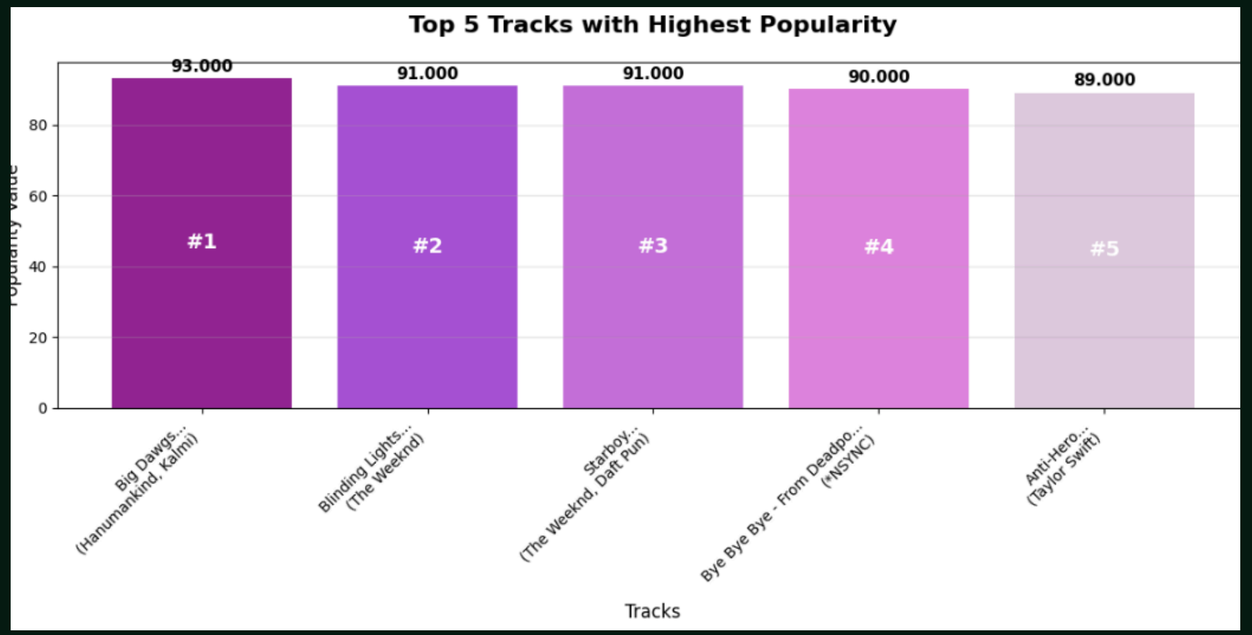
This bar chart showcases the top five tracks with the highest energy levels. All five songs have energy values between 0.999 and 1.000—the maximum possible—indicating they are extremely energetic and likely characterized by fast-paced, intense musical elements.

Bar chart with Highest Trempo



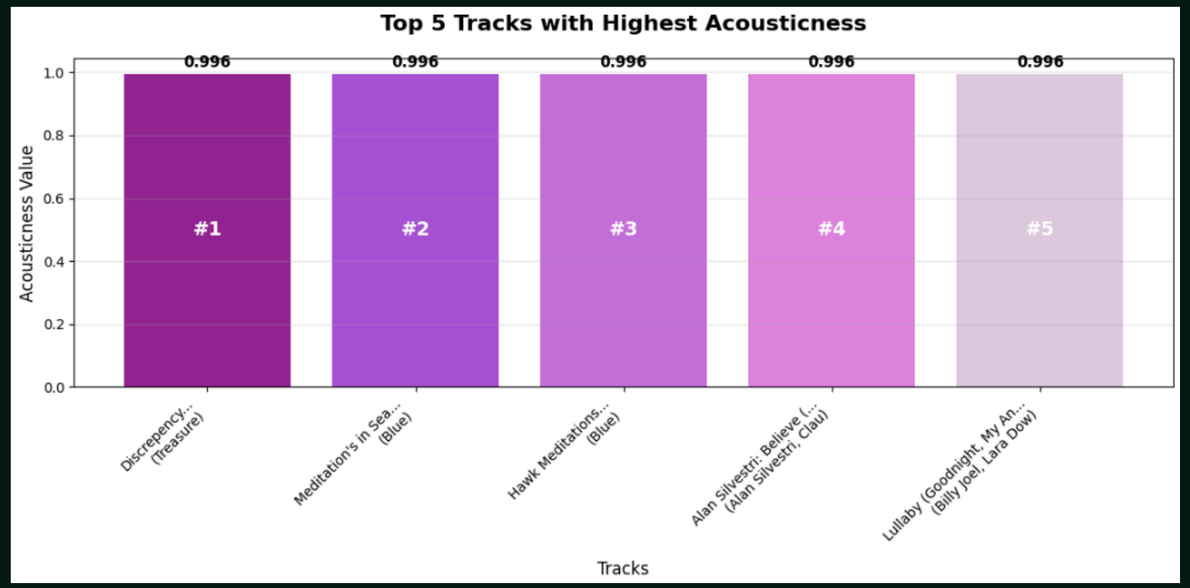
This bar chart presents the top five tracks with the highest tempo values, ranging from approximately 229.9 to 239.97 BPM. These values indicate that all the tracks are exceptionally fast-paced, reflecting a high-energy rhythm and intensity.

Popularity



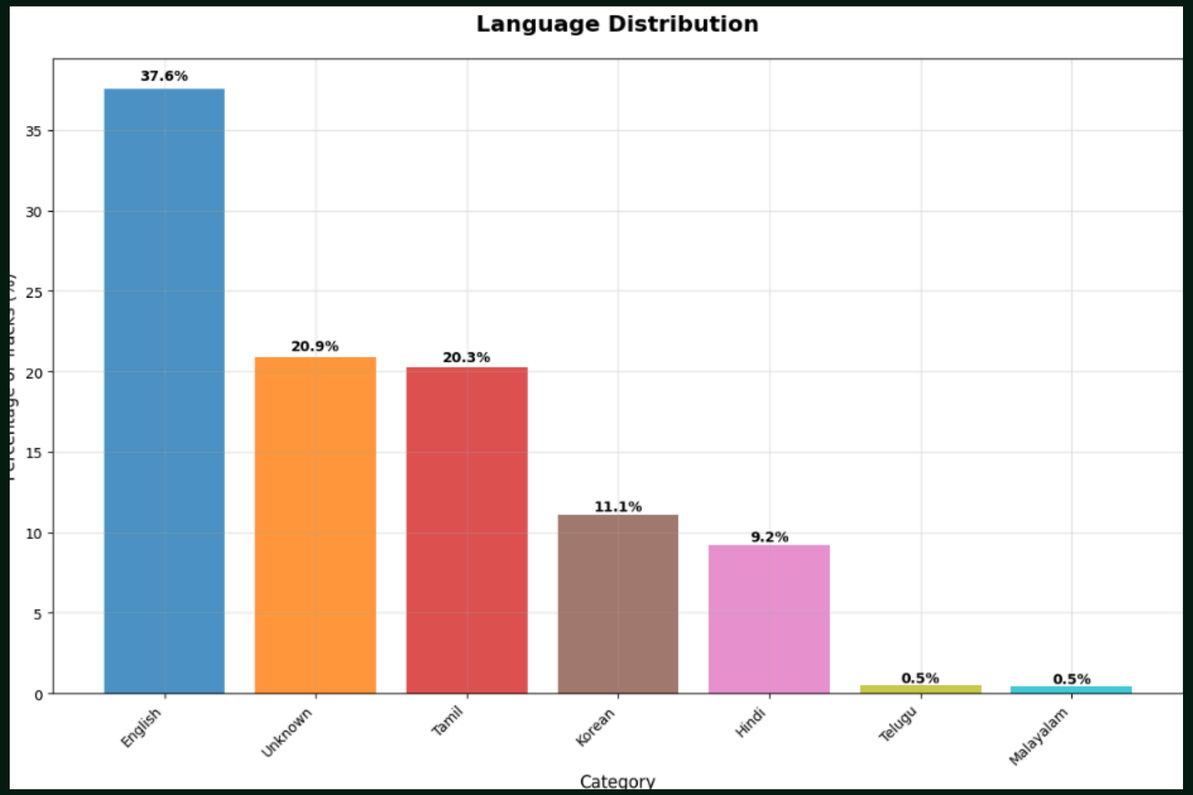
The top five tracks are all highly popular, with popularity scores ranging from 89.000 to 93.000.
Among them, “Big Energy” by Latto, Mariah Carey, and DJ Khaled stands out as the most popular track, achieving the highest score of 93.000.

Acousticness



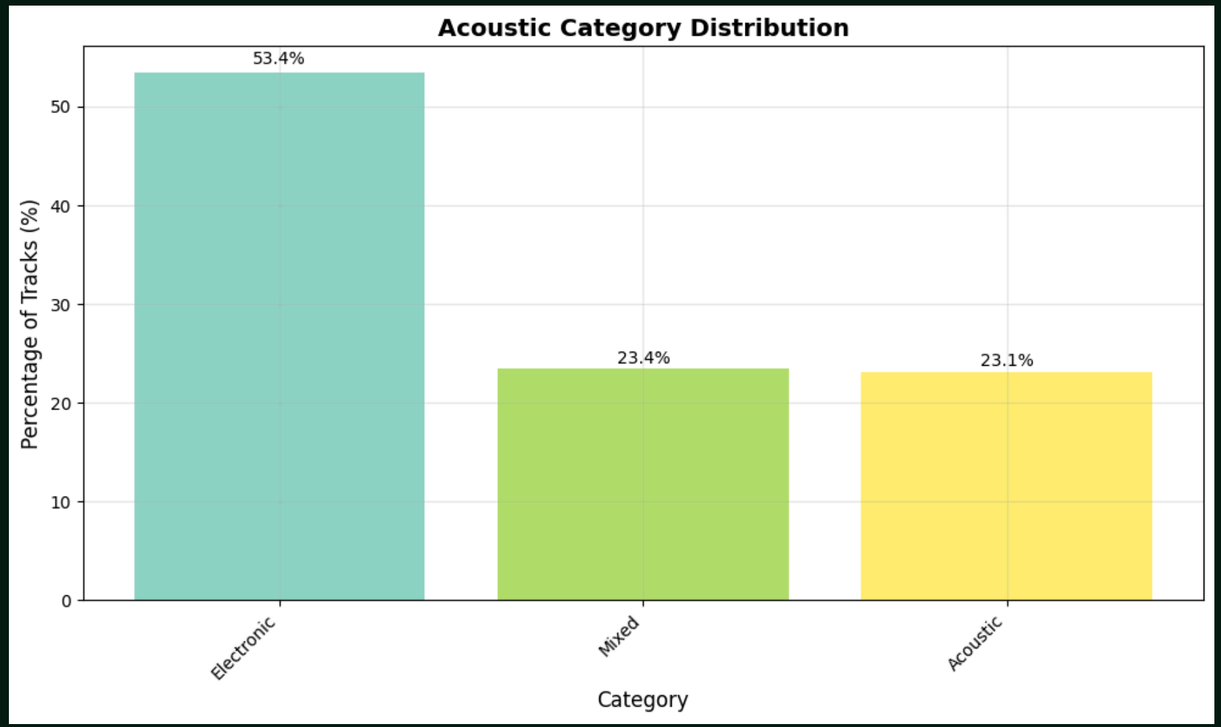
All five top tracks are nearly perfectly acoustic, each scoring 0.996.
This indicates that they have minimal to no electronic or synthesized elements, showcasing a strong emphasis on natural instrumentation.

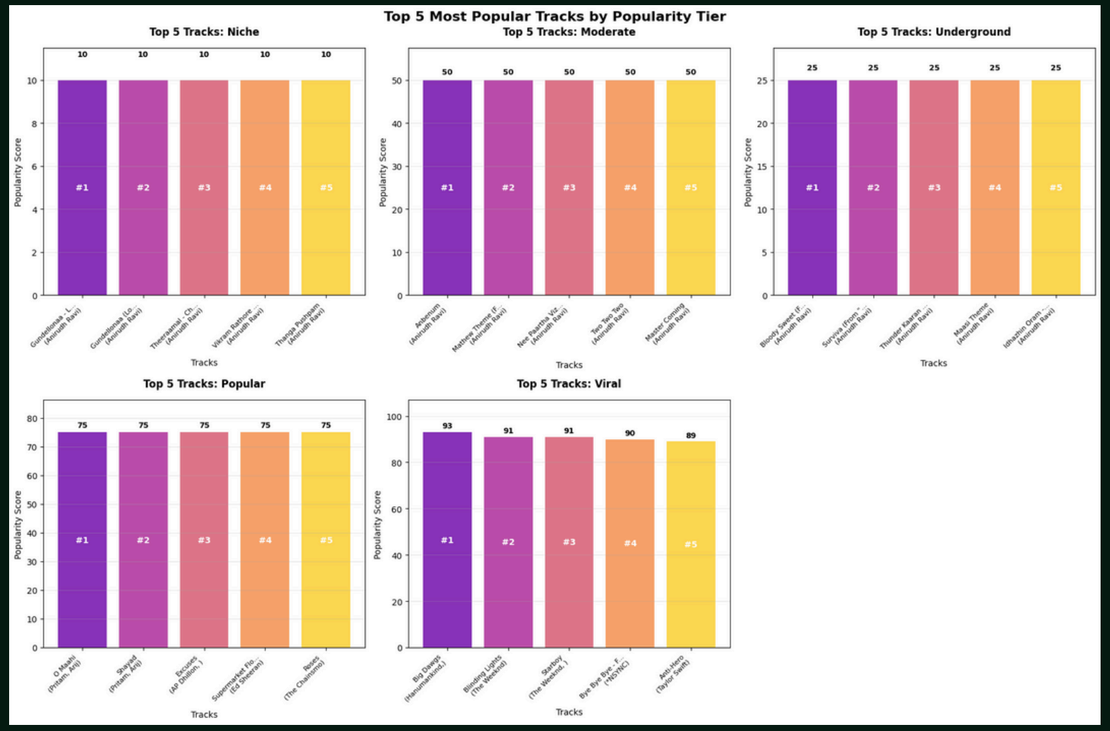
Donut charts for key categorical variables



- The Language Distribution shows that:
- English tracks dominate with 37.6%, followed by Unknown tracks at 20.9%, and Tamil at 10.3%.
- Combined, English and Unknown make up about 58.5% of all tracks.
- Korean (11.1%) and Hindi (9.2%) are the next major language groups.
- Telugu and Malayalam contribute minimally, each representing only 0.5% of the dataset.

- Rank #1: “Acrobatic Spasmolin” by Harushi Muratori Rusa has the highest loudness at −0.176 dB, nearly reaching the 0 dB digital maximum.
- The top three tracks (“Acrobatic Spasmolin,” “Dunkit,” and “Yanu”) have very similar loudness levels, ranging from −0.176 dB to −0.320 dB.
- The least loud among the top five is “I Hoo Crazy Abo” by Nicu Dane, with a loudness of −1.638 dB.



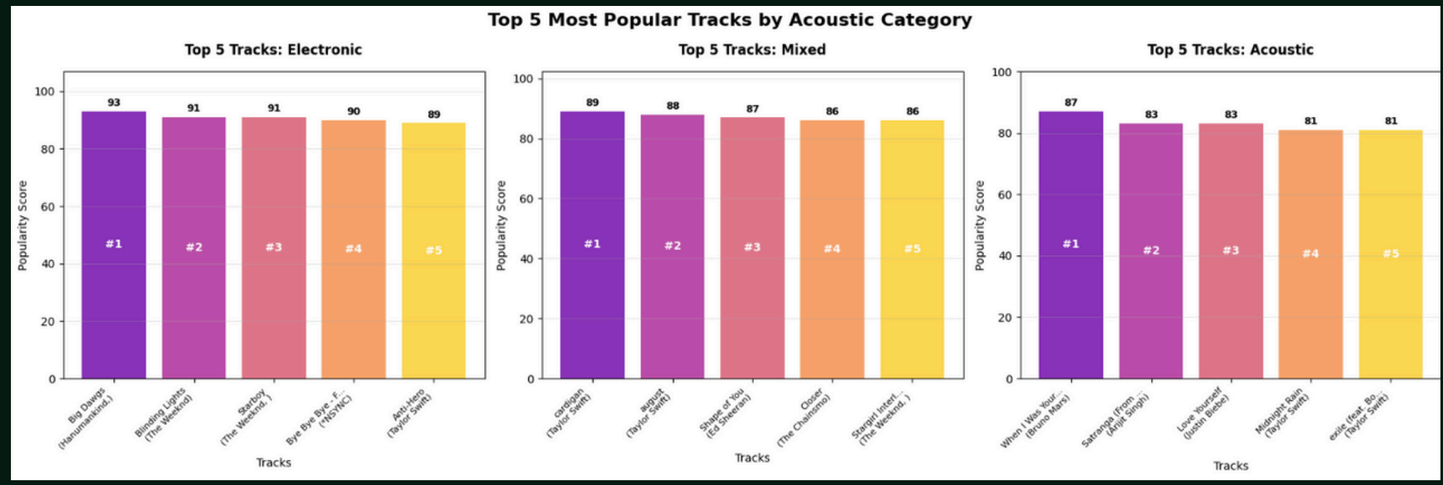


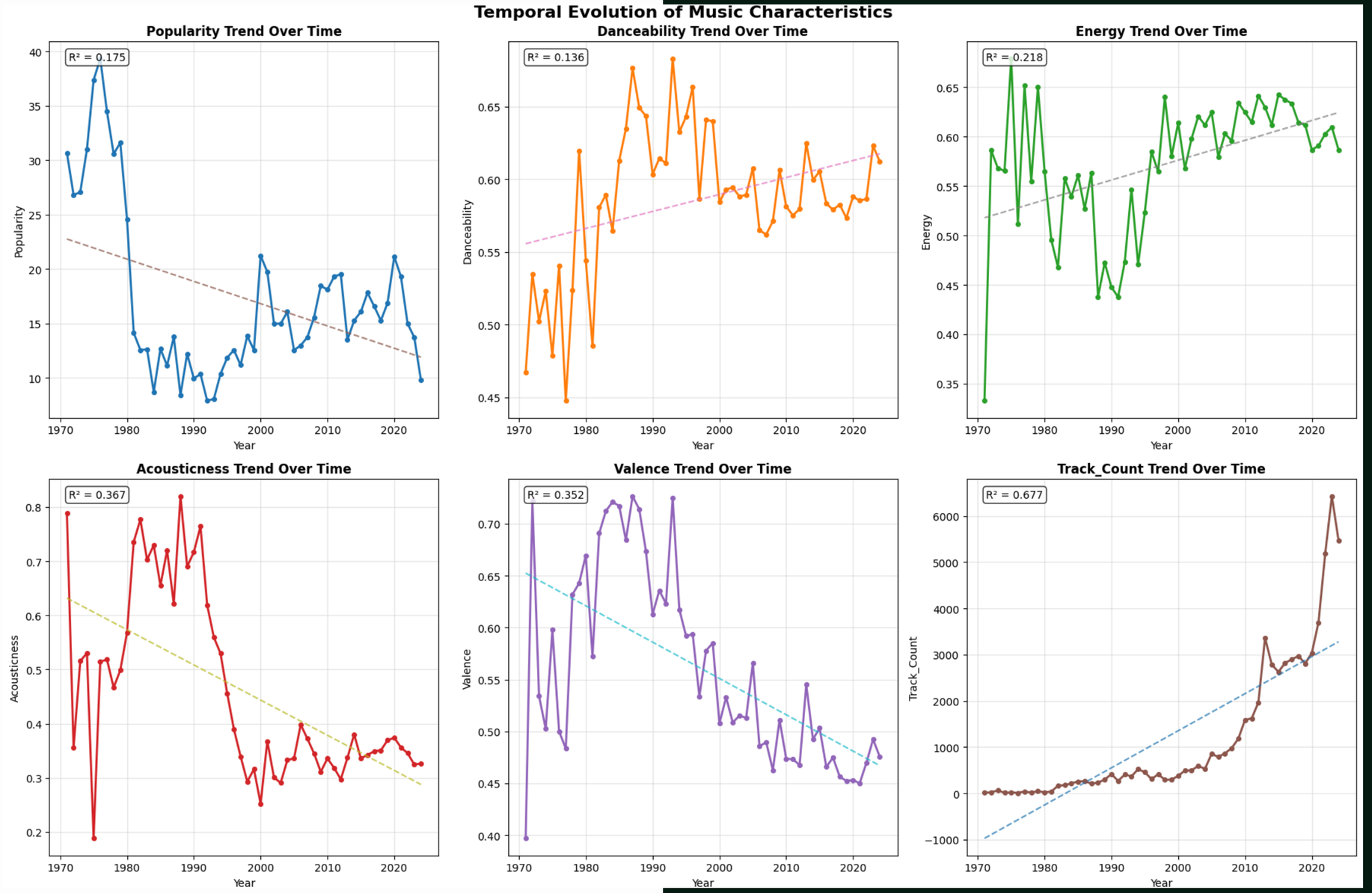
Popularity Tier Analysis:

- Niche Tier's High Score: The Niche tier includes the highest-scoring track overall, with a popularity score of 94 — an unexpected result for a category typically associated with limited appeal.
- Consistent High Popularity: All five tiers feature strong performers, with the lowest top score being 86, indicating consistently high popularity across categories.
- Viral and Popular Tiers: The Popular tier peaks at 92, while the Viral tier follows closely with a 90.
- Overall Insight: Even less mainstream tiers like Niche and Underground contain highly popular tracks, suggesting that strong listener appeal can exist across all tier classifications.

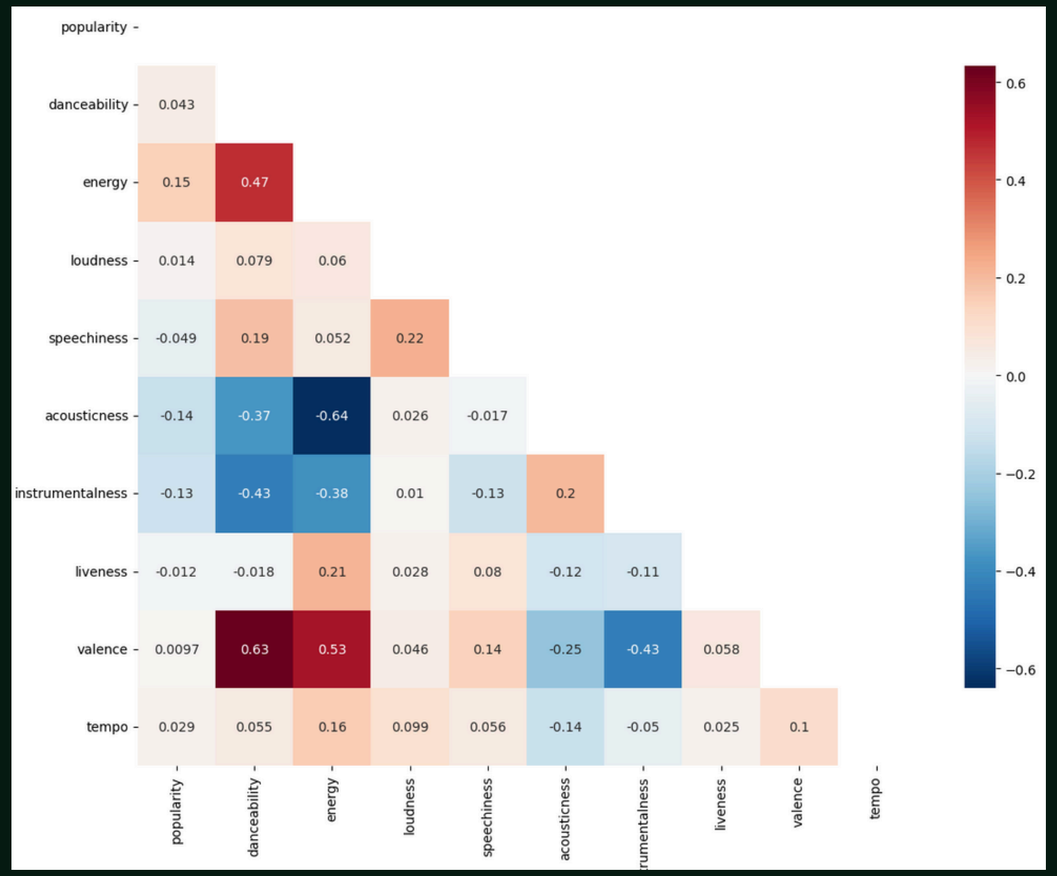
Popularity by Acoustic Category:

- Electronic: Leads in popularity, with the highest track scoring 94 and four out of five tracks scoring 90 or above.
- Mixed and Acoustic: Top tracks also perform strongly, peaking at 89 for Mixed and 87 for Acoustic.
- Tight Clustering: Across all three categories, top track popularity scores are closely grouped, ranging from 94 to 85.
- Overall Insight: Although Electronic tracks slightly edge out in peak popularity, the most popular songs remain consistently high-scoring across Mixed and Acoustic genres as well.

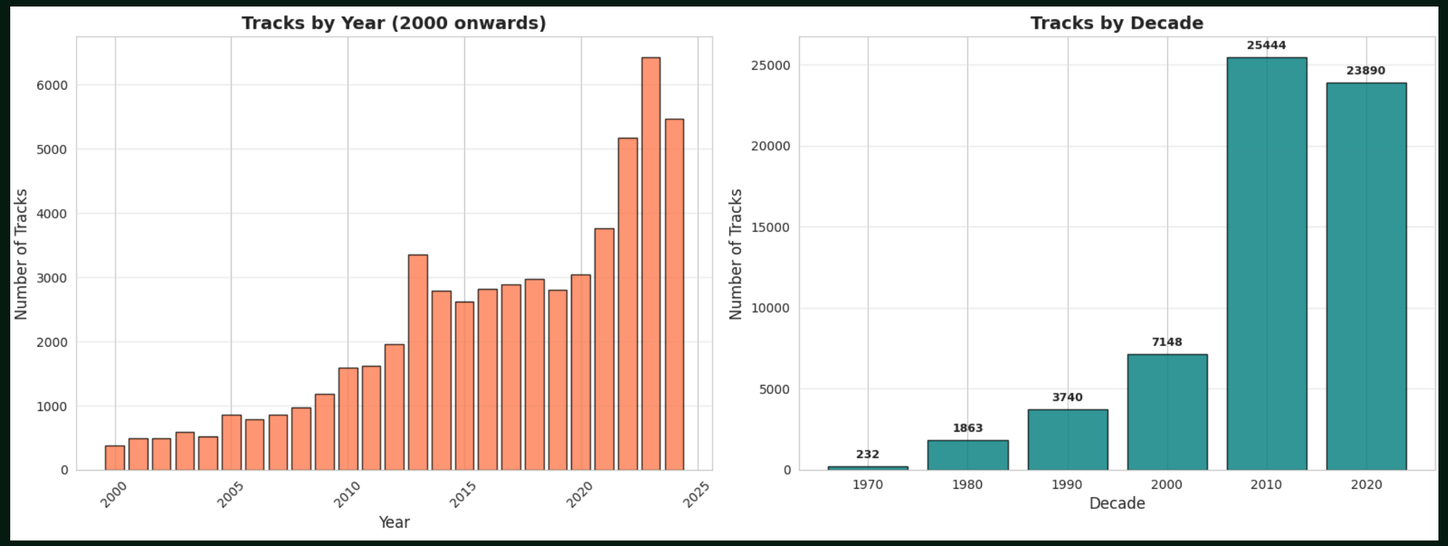




Temporal Analysis



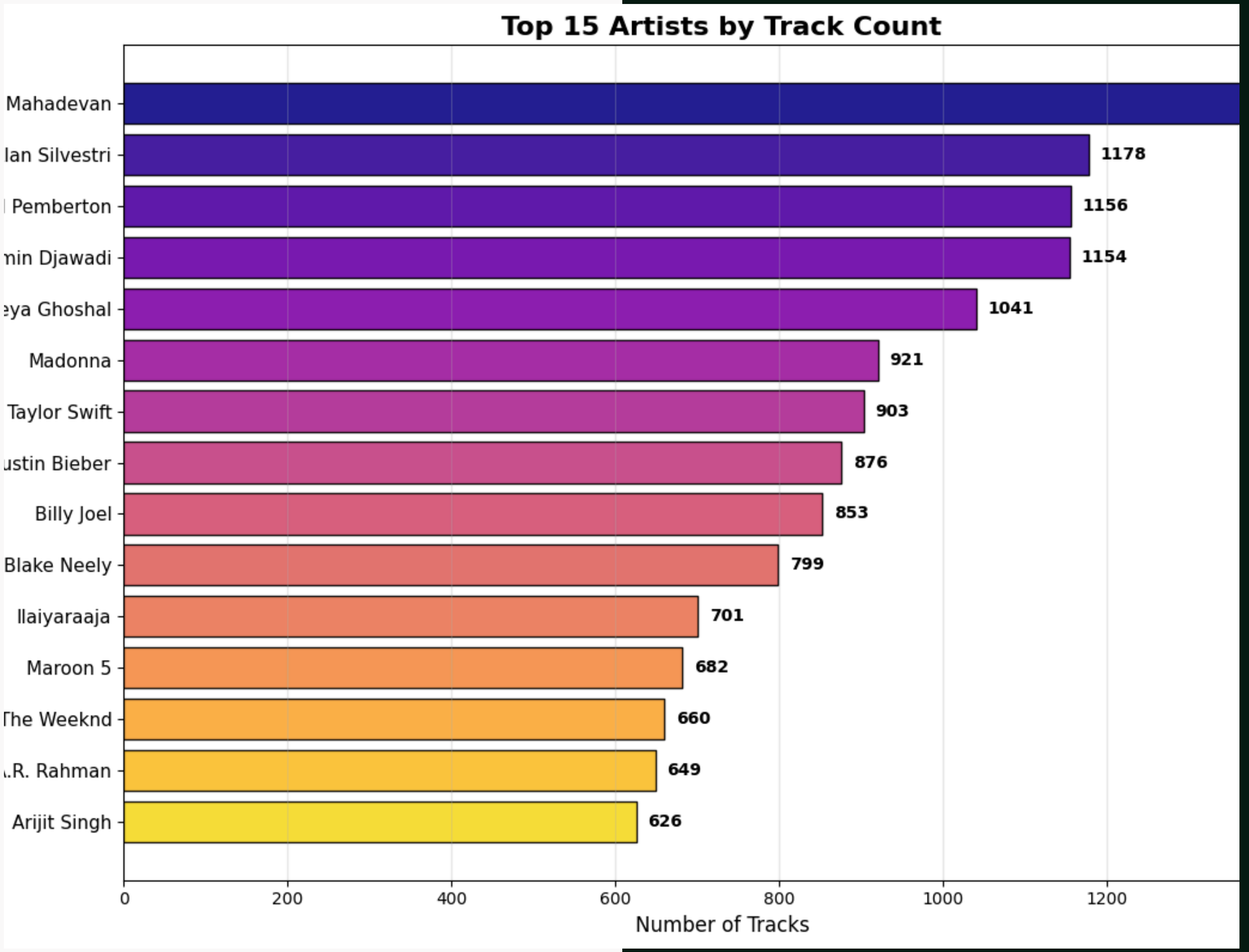
Bivariate Analysis and Correlations



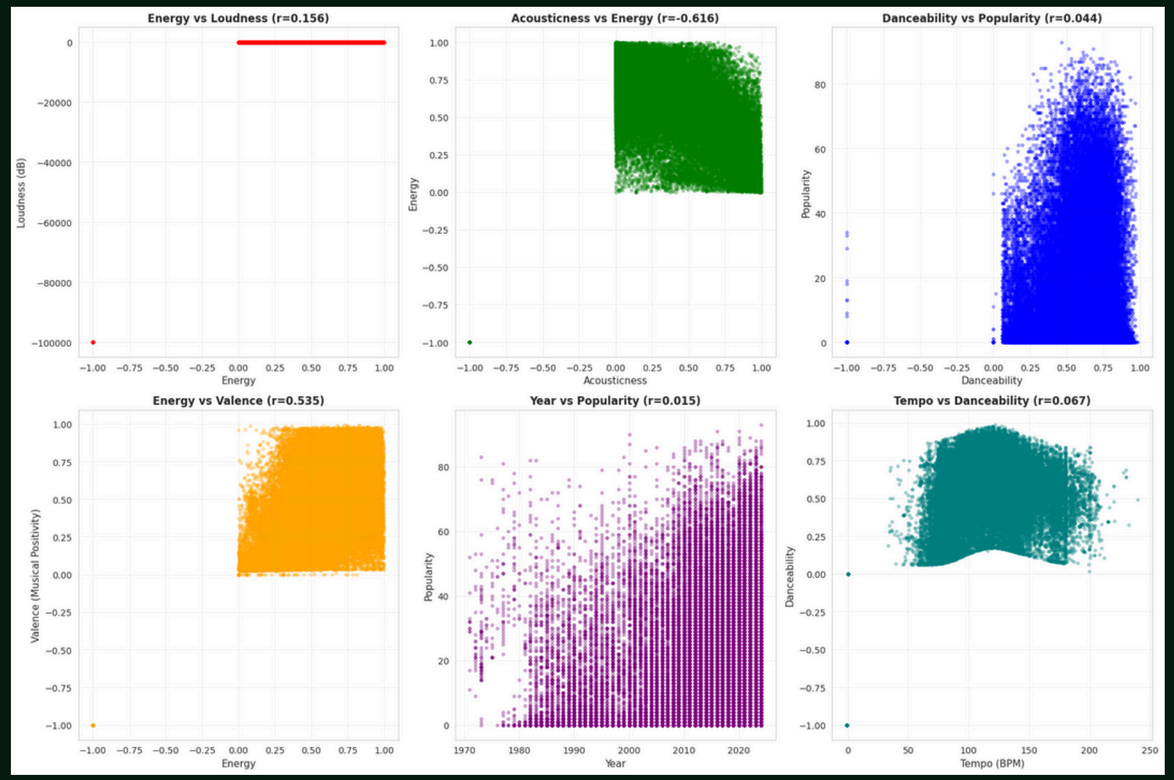
CATEGORICAL ANALYSIS - YEAR DISTRIBUTION

The dataset is heavily skewed toward recent decades, with the 2010s and 2020s accounting for over 75% of tracks.

Focusing on yearly trends since 2000, there is a rapid and accelerating increase in track releases, culminating in a sharp peak around 2023.



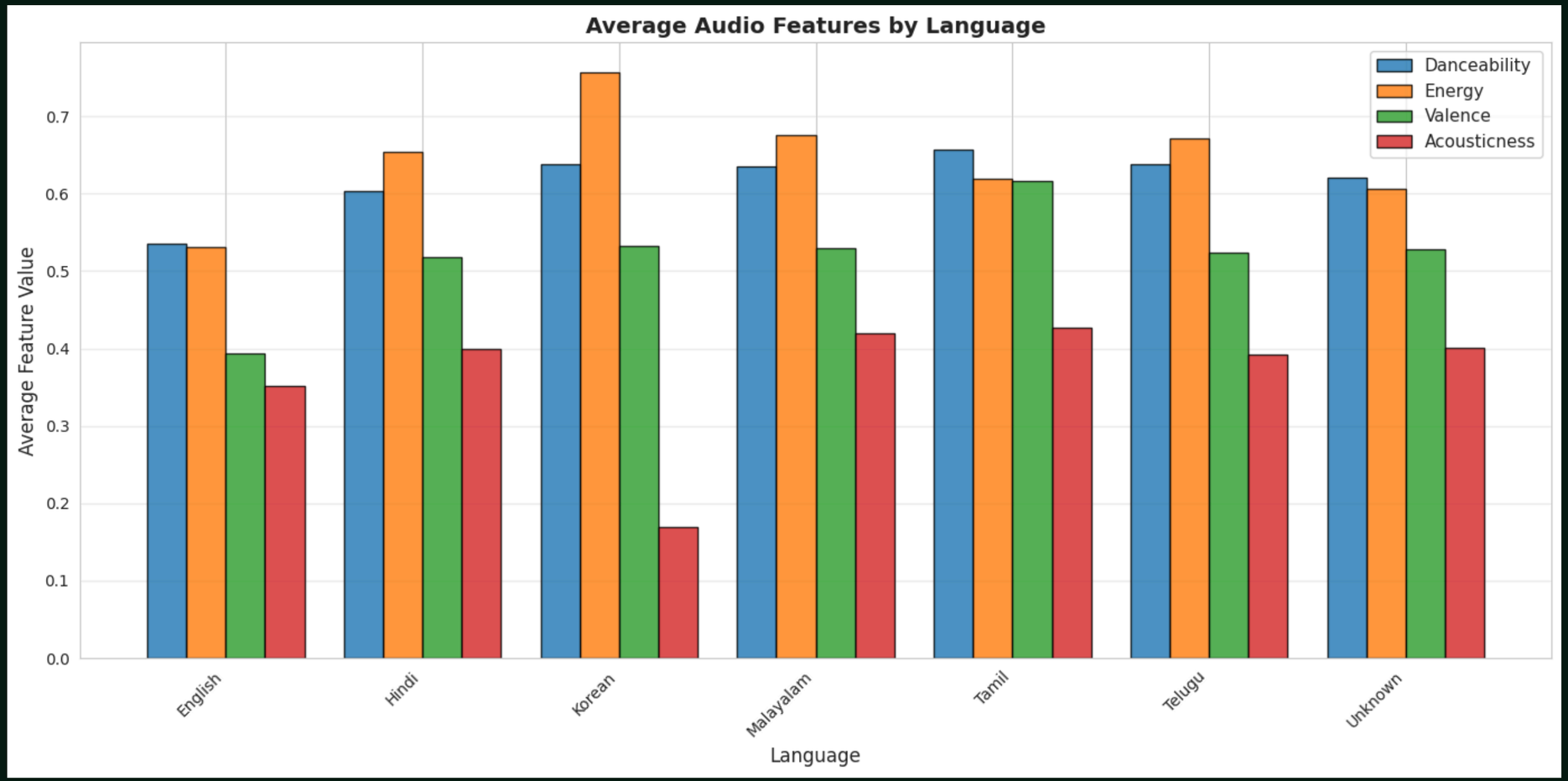
CATEGORICAL ANALYSIS - TOP ARTISTS



Bivariate Scatter Plots - Numerical Relationships

- The plots confirm a strong positive correlation between Energy and Valence (r=0.535), meaning that tracks with higher musical energy are also typically perceived as more positive or happy.
- There is a clear strong negative correlation between Acousticness and Energy (r=−0.616), which visually presents as two separate clusters of data—acoustic tracks are almost exclusively low in energy.
- Features like Danceability and Popularity (r=0.044) and Year and Popularity (r=0.015) show very weak correlations, indicating that a track's age or its danceability score are poor predictors of its popularity.

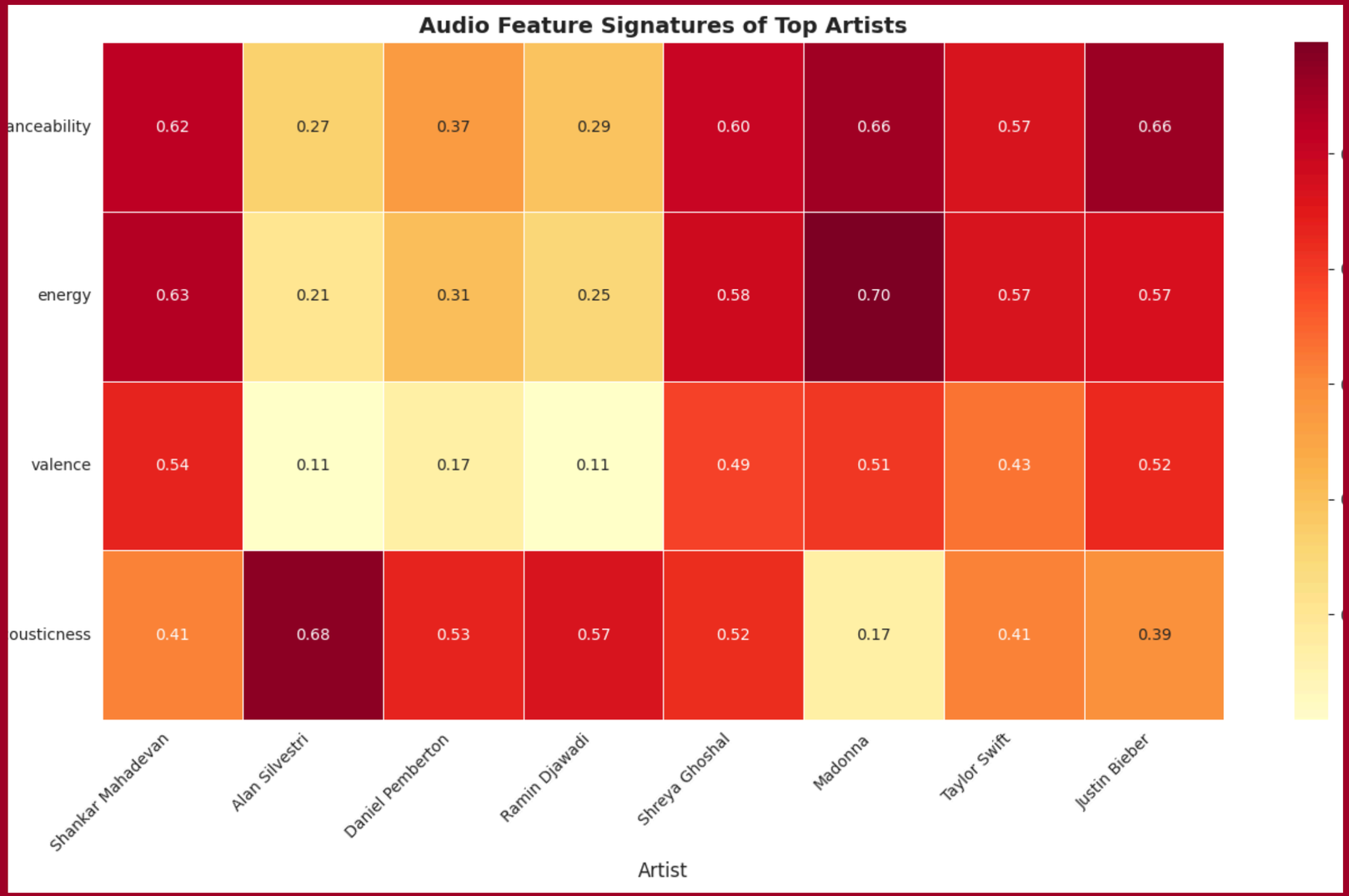
Bivariate Scatter Plots - Numerical Relationships



TOP ARTISTS: AUDIO FEATURE SIGNATURES

Madonna exhibits the highest levels of both Danceability (0.66) and Energy (0.70), suggesting her tracks are the most consistently upbeat and suitable for dancing among the top artists shown.

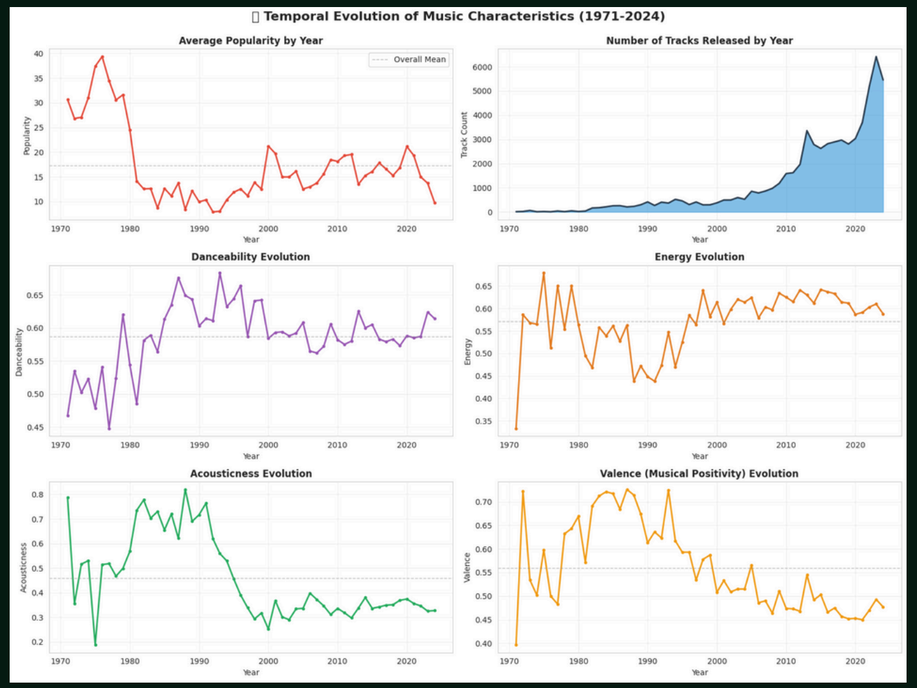
Alan Silvestri stands out with an exceptionally high Acousticness (0.68), paired with the lowest scores in Danceability (0.27), Energy (0.21), and Valence (0.11), indicating his music is predominantly non-energetic, acoustic, and likely cinematic/instrumental.

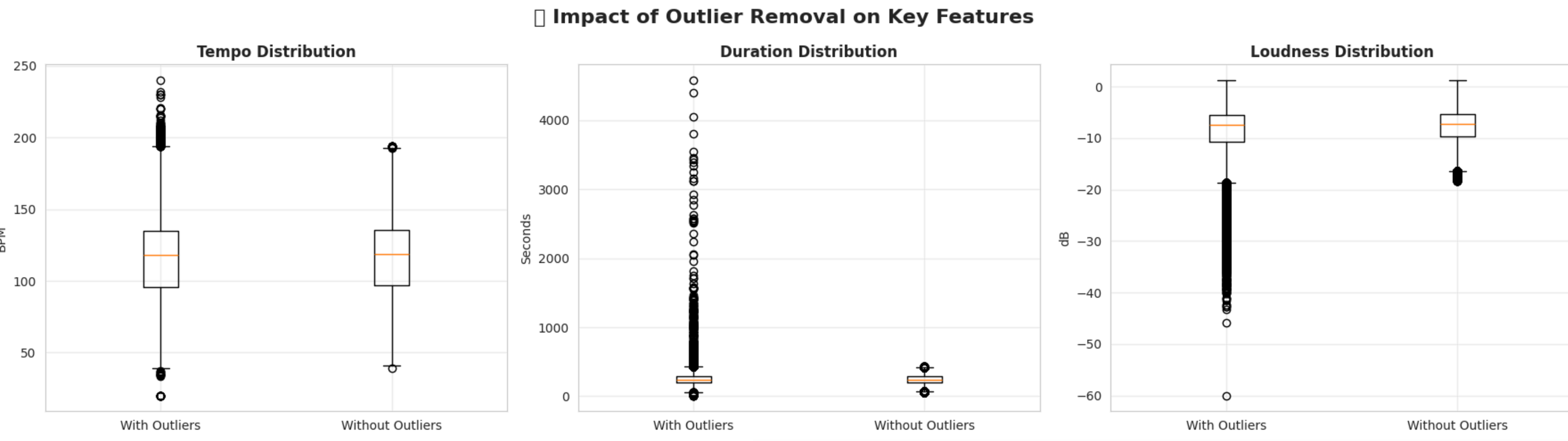


Time-Based Analysis: Trends Over Years

Temporal Trends in Music:

- Track Releases: The number of tracks released per year has grown exponentially since the early 2000s, with a major surge from 2010 to 2024.
- Popularity & Valence: Both average popularity and musical positivity (valence) peaked in the 1970s–1980s but have declined sharply since, remaining below the overall mean for most years after 1990.
- Acousticness & Energy: Acousticness has dropped dramatically since around 1990, while Energy has generally increased, staying above its overall mean for most of the 21st century.
- Insight: Modern tracks are more energetic and less acoustic, yet they don’t necessarily achieve the same popularity or musical positivity as earlier decades.



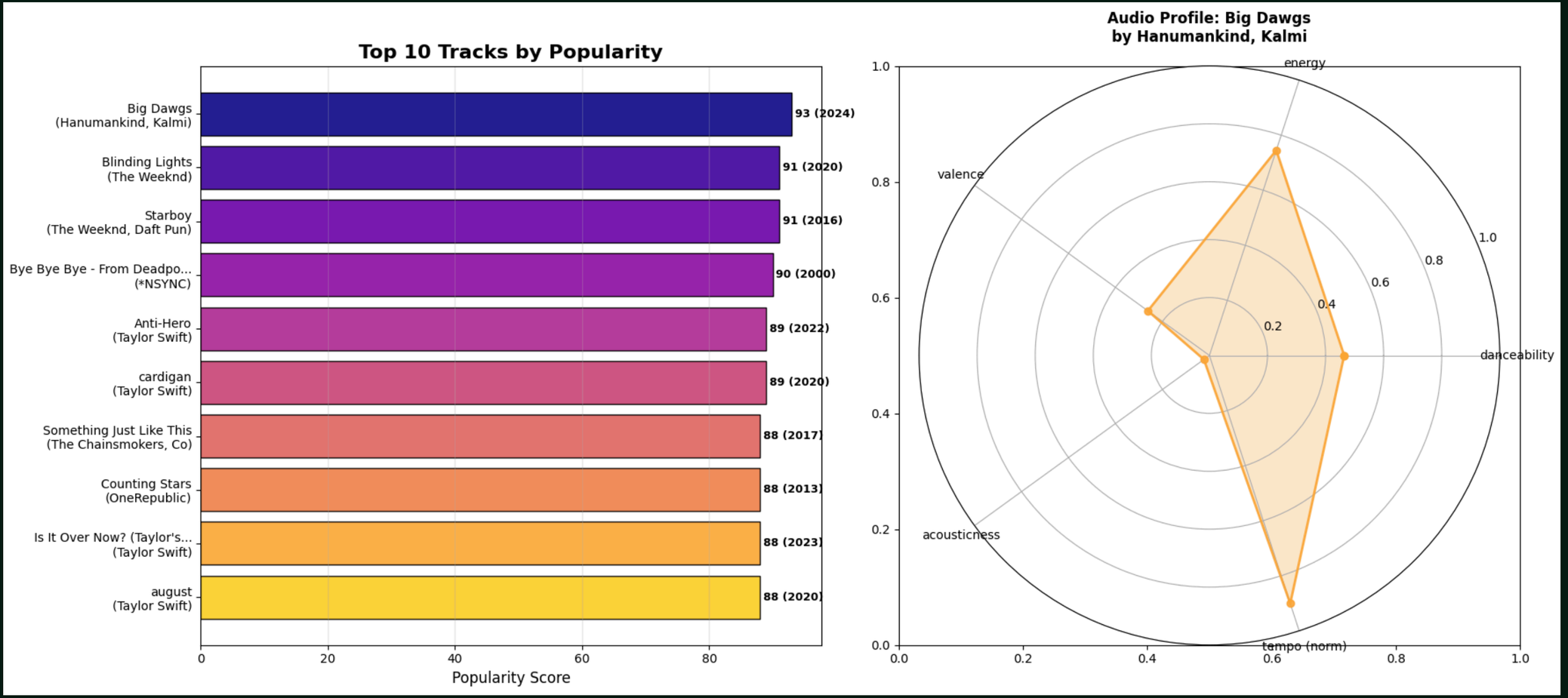


OUTLIER ANALYSIS & TREATMENT

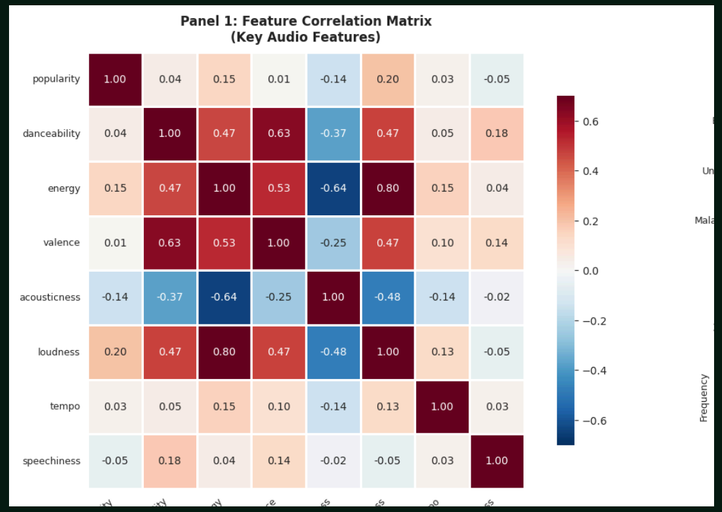
Distribution Insights for Key Features:

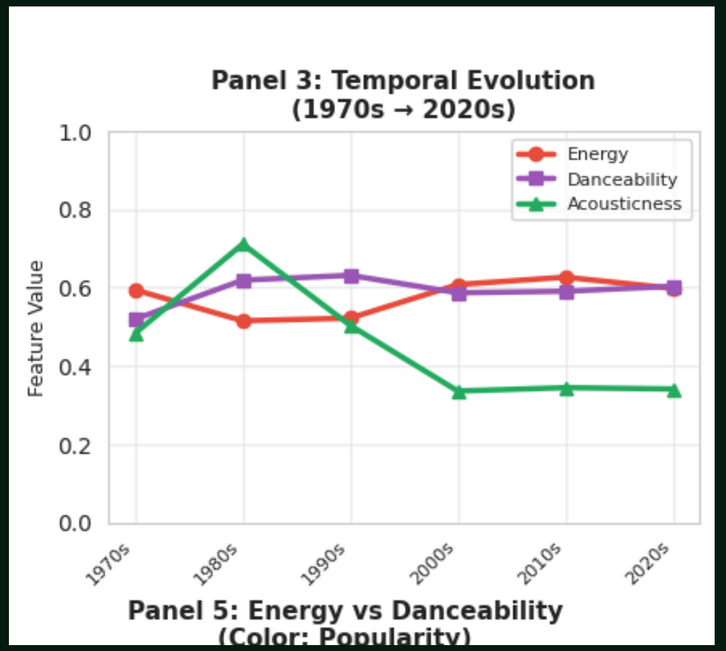
- Popularity & Duration_ms: Both features show a high number of outliers (Popularity: 1,115; Duration_ms: 2,272), with many tracks well above the upper quartile, indicating long, thin tails in their distributions.
- Energy: The distribution is nearly perfectly symmetrical with no outliers, centered around a median of 0.64, showing a tight, uniform spread of values.
- Tempo: Features relatively few outliers (412, 0.7%) and a narrow interquartile range (IQR: 156.51), indicating that most tracks have similar tempos and a compact distribution.
- Insight: While popularity and duration vary widely, features like energy and tempo are more consistent, reflecting standardization in modern track characteristics.

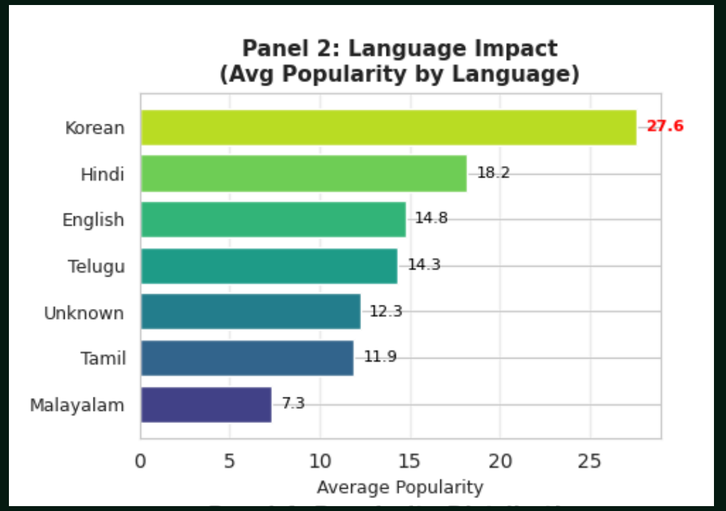
Music Popularity Analysis & Audio Profile



FINAL COMPREHENSIVE REPORT DASHBOARD







Spotify Data Analysis: Comprehensive Report Summary

Data Composition and Distribution:
The dataset is heavily focused on modern music, with over 79% of tracks from the 2010s and 2020s. High-energy songs make up 55.5% of the data. Korean tracks show the highest average popularity (27.6), surpassing Hindi (18.2) and English (14.8) songs.

Key Feature Correlations:
Energy and Valence are strongly correlated (0.64), while Acousticness and Energy show a strong negative link (-0.58). Popularity has only weak positive ties with Danceability and Energy.

Temporal Evolution and Insights:
Over time, Acousticness has declined, while Danceability and Energy have risen. Despite this, Korean music stands out in popularity, suggesting cultural and language factors influence success more than audio traits alone.

CONCLUSION

- The dataset provided valuable insights into trends among Spotify tracks across different genres and artists.
- Key audio features like danceability, energy, and tempo significantly influence a song’s popularity.
- Popular songs often exhibit similar audio feature patterns, allowing identification of characteristics common to hit tracks.
- Artists who maintain a consistent audio style tend to keep their listeners more engaged.
- Data visualization helped in effectively comparing features and revealing correlations between them.
- The analysis highlights how data science techniques can extract meaningful insights from extensive music datasets.
- Overall, the project demonstrates that Spotify data can be leveraged to understand musical trends and predict song popularity.

Thank You